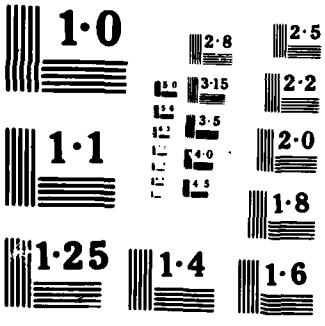
AD-A157 294 COMPUTERIZED ADAPTIVE MEASUREMENT OF ACHIEVEMENT AND ABILITY(U) MINNESOTA UNIV MINNEAPOLIS COMPUTERIZED ADAPTIVE TESTING LAB D J WEISS JUN 85 N00014-79-C-0172 1/1

UNCLASSIFIED F/G 9/2 NL

END
DATE FILMED
11-85
DTIC


1·0
1·1
1·25
1·4
1·6
1·8
2·0
2·2
2·5
2·8
3·15
3·5
4·0
4·5

Final Report
Computerized Adaptive Measurement of Achievement and Ability

David J. Weiss

June 1985

COMPUTERIZED ADAPTIVE TESTING LABORATORY
DEPARMENT OF PSYCHOLOGY
UNIVERSITY OF MINNESOTA
MINNEAPOLIS MN 55455

FINAL REPORT OF PROJECT NR150-433, N00014-79-C-0172
SUPPORTED BY THE
OFFICE OF NAVAL RESEARCH
AIR FORCE HUMAN RESOURCES LABORATORY
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
ARMY RESEARCH INSTITUTE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER	2. GOVT ACCESSION NO. AD-A157294	3. RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle) Final Report: Computerized Adaptive Measurement of Achievement and Ability	5. TYPE OF REPORT & PERIOD COVERED Final Report February 1979 - April 1983 6. PERFORMING ORG. REPORT NUMBER
7 AUTHOR(s) David J. Weiss	8. CONTRACT OR GRANT NUMBER(s) N00014-79-C-0172
9 PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Minnesota Minneapolis, MN 55455	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS PE: 615534 Proj: RR042-04 TA: RR042-04-01 WU: NR150-433
11 CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, VA 22217	12. REPORT DATE June 1985 13. NUMBER OF PAGES 20
14 MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approval for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18 SUPPLEMENTARY NOTES

This research was supported by funds from the Office of Naval Research, Air Force Human Resources Laboratory, Air Force Office of Scientific Research, and Army Research Institute, and monitored by the Office of Naval Research.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

The research program's objectives are described, and the research approach is summarized and related to the ten technical reports and other project publications. Thirteen major research findings are presented. Abstracts of the ten technical reports are also included.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

CONTENTS

FINAL REPORT

COMPUTERIZED ADAPTIVE MEASUREMENT OF ACHIEVEMENT AND ABILITY

Objectives

This research program was designed to investigate the applications of item response theory (IRT) and computerized adaptive testing to the unique problems of the measurement of ability and the measurement of achievement. Specific objectives relevant to these two areas were as follows:

Adaptive Achievement Testing

1) To study the relative efficiency of various approaches to intersubtest branching in achievement test batteries.

2) To investigate the dimensionality of measured achievement over time.

3) To study the applicability of IRT models to the problem of mastery testing and to compare models for adaptive mastery testing with other approaches to the improvement of mastery decisions and/or reduction in test length in mastery testing.

4) To explicate the concept of Adaptive Self-Referenced Testing and to examine its applicability to the achievement testing problem.

Adaptive Ability Testing

5) To evaluate the performance of adaptive testing strategies under conditions which more reasonably represent the conditions under which these strategies might be used, and to examine the performance of adaptive testing strategies in live testing.

6) To evaluate the utility for adaptive testing of response modes and test item formats usable in adaptive ability testing.

Research in pursuance of these objectives began in February 1979 and continued through April 1983.

Approach

The research utilized a combination of monte carlo simulation studies and live-testing studies.

Adaptive Achievement Testing

Intersubtest branching. Intersubtest branching is an approach to the utilization of adaptive testing methodologies in a multidimensional item pool. In intersubtest branching, IRT item parameters are estimated separately for each subtest of a multisubtest battery. Using any of a number of adaptive testing strategies, adaptive testing occurs within the subtest based on appropriate item selection rules and a test termination criterion appropriate for the purpose of

testing. Upon completion of a subtest in the test battery, the final trait level estimate ($\hat{\theta}$) is then used as an entry point to begin testing in a subsequent subtest in the battery. As originally proposed, subtests in a battery are ordered by the magnitudes of the squared multiple correlations of each subtest with all other subtests in the battery. In this way, the entry points for adaptive testing in each subtest utilize the information available in the tests in the test battery that were most highly correlated with it, which should shorten the adaptive tests for later subtests in the battery as much as possible.

Intersubtest adaptive branching was studied by real-data simulation in Research Report 79-6, and by monte carlo simulation in Research Report 80-4. The study reported in Research Report 79-6 used data from conventionally-administered tests which were analyzed as if they had been administered as an adaptive test, and the intersubtest branching strategy was applied to these data. This study was designed to separate the effects of the adaptive intrasubtest item selection procedure from those effects due to intersubtest branching. The study also (1) allowed evaluation of the effects of different intrasubtest termination criteria, (2) investigated the effect of taking into account errors of measurement in the multiple regression procedure used to determine test entry points, and (3) investigated the stability of the regression equations in cross-validation.

Other aspects of the intersubtest branching strategy when applied to an achievement test battery were investigated by monte carlo simulation in Research Report 81-4. Questions of interest in this study included (1) the effects of varying subtest order, (2) the utilization of different subtest termination criteria, and (3) the effect of variable versus fixed entry on the psychometric properties of the intersubtest branching strategy. Dependent variables included (1) reductions in test length, (2) effect on test information, and (3) correlations between achievement estimates and true achievement levels. The study design also permitted separation of the effects of intrasubtest and intersubtest adaptive branching.

The dimensionality of measured achievement over time. The effects of instruction on measured achievement are usually measured at a single point in time. That is, some instruction is given to an individual and at the end of the period of instruction an achievement test is used to determine whether the individual has reached an appropriate level of achievement. On the basis of such information, aggregated across individuals, decisions are frequently made about the adequacy of instructional programs, or about the impact (or lack thereof) of instruction on a specific individual.

A more powerful approach to the measurement of achievement would involve the use of pretests and posttests to determine if any change has occurred in measured achievement over time. Using change scores, however, implies that the variable being measured is the same at pretest as it is at posttest. There has been very little empirical data available concerning this issue.

Research Report 81-5 was designed to investigate the question of whether the achievement factor identified at pretest in an achievement test is the same factor identified at posttest. Two studies utilized data on groups of college students from measured achievement in mathematics classes and biology classes.

Achievement test item responses were factor analyzed prior to instruction, and again at the end of instruction. In addition, mean differences in test scores at pretest and posttest were analyzed. Factors obtained at pretest were compared with those obtained at posttest to determine if the same factor was found prior to and after instruction.

Kingsbury (1984) directly examined the characteristics of change scores derived from adaptive and conventional tests. This study utilized data from college-level biology examinations. Both adaptive and conventional tests were administered in a complex design to groups of students in such a way that reliabilities of the change scores could be determined separately for the two types of tests in a number of different homogeneous content areas covered in the course. The question raised by this study was based on hypotheses that the more precise achievement level estimates resulting from adaptive testing should also result in more reliable change scores in comparison to those from conventional testing. Also studied was the effect of variable- versus fixed-length test termination on the adaptive tests.

Adaptive Mastery Testing. Adaptive mastery testing (AMT) combines IRT and adaptive testing into an efficient strategy for making mastery or classification decisions. In this procedure, items used to make a mastery decision are selected by an IRT maximum information adaptive testing strategy. Item responses are scored using a Bayesian θ estimation procedure, and a confidence or credibility interval is computed for the θ estimate. The confidence interval around the estimate is then compared to a mastery cutoff score, which is also expressed on the θ metric. A mastery decision is determined on the basis of whether the credibility interval overlaps with the mastery criterion level, and on which side of the mastery cutoff score the individual's θ estimate falls.

Both monte carlo simulation and live testing were used to investigate characteristics of the AMT strategy and to compare it with other approaches for making mastery decisions. In Research Report 80-4 (also Kingsbury & Weiss, 1980a) the AMT procedure was compared to a conventionally-based mastery testing procedure and to a procedure based on Wald's sequential probability ratio test. The procedures were compared in terms of their efficiency, based on the test length required by the procedures to make a classification decision, on the validity of the decisions made by each procedure, and on the type of classifications made by each of the three testing procedures.

To examine the generality of the findings in live testing, in Research Report 81-3 the AMT procedure and a conventional test were administered to students in a biology class. Contrary to earlier studies which examined the AMT procedure, actual adaptive mastery tests were administered to one subgroup of students while the other received computer-administered conventional tests. The performance of the two testing strategies was evaluated in terms of a mastery criterion based on the students' final standing in the course, which was a combination of their performance on course examinations and laboratory grades.

Adaptive Self-Referenced Testing. Adaptive self-referenced testing (ASRT) is a combination of IRT and adaptive testing designed to permit the efficient measurement of changes in achievement levels due to exposure to instruction. This procedure is designed to measure individual changes in achievement in a

unidimensional item pool in a very efficient manner at a number of points of instruction. It is thus an appropriate conceptualization for tracking individual changes due to instruction at a number of points during a course, since it permits an instructor to evaluate an individual's performance on a minimum number of items at each of a number of testing occasions.

ASRT permits an instructor to measure a student early in a course, such as on the first day, and as frequently as is necessary during the course. Based on adaptive testing methodology, the data obtained from the Time 1 testing are used as the entry point to Time 2 adaptive test administration, and this process is followed for any number of test administrations. In addition, test termination at any point in time can be based on the standard error band associated with an individual's θ estimate at that point in time. ASRT is designed to simultaneously permit intraindividual measurement of change, norm-based measurement on the θ metric which can then be converted to the proportion-correct measurement if desired, and a mastery-based (criterion-referenced) achievement level estimate utilizing the procedures of AMT. While no research directly related to ASRT was done during the contract period, the method was described in some detail in Weiss & Kingsbury (1984). Both Research Report 81-5 and the Kingsbury (1984) study have implications for the use of ASRT and its future development.

Adaptive Ability Testing

Adaptive testing strategies. A major focus of this research program was on the evaluation of different approaches to computerized adaptive testing. While earlier projects were concerned primarily with evaluating the relative performance of adaptive and conventional testing strategies, in this project the focus was on the IRT-based strategies and on their performance under a variety of conditions. An overview of some aspects of project research is given by Weiss (1982).

The performance of a Bayesian adaptive testing strategy was reported in Research Report 83-2 (also Weiss & McBride, 1984). Owen's Bayesian adaptive testing strategy was examined in three studies which utilized an accurate prior θ estimate, a constant prior θ estimate with fixed test length, and a constant prior θ estimate with variable test length. The performance of the adaptive testing strategy was examined in terms of the bias and information of the θ estimates as a function of θ. Also examined was the mean number of items administered in the variable test length condition.

A major concern of the research was to evaluate the performance of adaptive testing strategies under conditions of increasing realisticness. Prior to these studies, all studies evaluating the performance of adaptive testing strategies did so under reasonably unrealistic conditions. While characteristics of the item pools varied in these earlier studies, the IRT item parameters used in these simulation studies were considered to be accurate. However, in real item pools, there is always some error associated with the item parameter estimates. Since adaptive testing is designed to select items on the basis of these item parameter estimates, it can be assumed that any degree of inaccuracy in the item parameter estimates will have detrimental effects on the performance of adaptive testing strategies.

Consequently, two studies were designed to investigate effects of errors in item parameter estimates on the performance of maximum information and Bayesian adaptive testing strategies. The first study (Crichton, 1981) assessed the effects of errors in item parameter estimates in the context of the three-parameter logistic model. Crichton compared the performance of the two IRT-based adaptive testing strategies--maximum information and Bayesian--with the stratified adaptive (stradaptive) strategy, on the hypothesis that the stradaptive strategy should be less sensitive to errors in the item parameter estimates. Her monte carlo simulation study varied test length from 5 to 30 items. Test length was then crossed with three levels of error in the discrimination ($\underline{a}$) parameter, four levels of error in estimates for the difficulty ($\underline{b}$) parameter, and two levels of error in the pseudo-guessing ($\underline{c}$) parameter. In addition to considering these effects for $\underline{a}$, $\underline{b}$, and $\underline{c}$ separately, two datasets examined the effects of joint errors in the $\underline{a}$, $\underline{b}$, and $\underline{c}$ parameters. Dependent variables conditional on θ included the bias, root mean square error, inaccuracy, and information in the θ estimates, and the correlation of θ and $\hat{\theta}$.

Mattson (1983) also examined the performance of adaptive testing strategies under conditions of error in item parameter estimates, using monte carlo simulation. Mattson extended the Crichton study by studying similar effects in the one- and two-parameter logistic models, in addition to the three-parameter model. Whereas Crichton limited her trait level estimation to maximum likelihood scoring of the response vectors, Mattson also included Bayesian scoring of the maximum information and Bayesian adaptive tests. In addition, Mattson allowed the level of correlation between the $\underline{a}$ and $\underline{b}$ parameters to vary at four different levels, as well as examining the uncorrelated condition used by Crichton. Similar to Crichton, Mattson also varied test length from 10 to 30 items. Finally, Mattson allowed errors in the $\underline{a}$ parameter to vary at two levels, examined four levels of error in $\underline{b}$, and one level of error in $\underline{c}$. All conditions were crossed with each other. Mattson's dependent variables were the same as those studied by Crichton.

A second factor that can affect the performance of adaptive testing strategies in a realistic item pool is the dimensionality of the item pool. Since all IRT models assume a unidimensional item pool, deviations from unidimensionality would be expected to affect the performance of adaptive testing strategies in real item pools, which are rarely (if ever) strictly unidimensional. As a result Suhadolnik and Weiss (1985) examined the robustness of adaptive testing to multidimensionality.

In this study, the maximum information adaptive testing strategy using maximum liklihood scoring was applied to datasets varying from strictly unidimensional to four-factor datasets that reflected the structure of the most multidimensional subtest of the Armed Service Vocational Aptitude Battery. Between these extremes were two- and three-factor datasets in which the second and third factors accounted for varying proportions of variance in comparison to the first factor, thus simulating item structures varying from very little multidimensionality, to a very high degree of multidimensionality. A total of 45 data structures was examined.

To evaluate the effects of multidimensionality, dichotomous item responses were simulated from the specified multidimensional structures. These item re-

sponses were then treated as if they were derived from a unidimensonal model, and adaptive testing was implemented using the item response vectors. To evaluate the performance of the maximum information adaptive testing strategy under multidimensionality, the conditional bias, inaccuracy, and root mean square error of the θ estimates was computed relative to the true first factor θ from the multidimensional structure.

Response modes, test item formats, and effects of test administration variables. The administration of ability tests by interactive computers allows the use of item types that transcend the typical dichotomously-scored multiple-choice test item. Research Report 83-3 examined aspects of a probabilistic response mode used in conjunction with the typical multiple-choice item format. This response mode was chosen as one means of extracting additional information from a multiple-choice item, rather than simply requiring a choice of a single response alternative.

A major problem with probabilistic responding to multiple-choice items in conventional paper-and-pencil test administration is that examinees do not always follow the instructions carefully so that the probabilities they assign to the item responses does not always sum to 1.00. As a consequence, large amounts of data might be lost for a given examinee. When multiple-choice items are answered in a probabilistic mode on a computer terminal, however, the validity of the distribution of the probabilities can be checked immediately for each individual's responses to each test item, and invalid responses can be adjusted until they meet the appropriate criteria.

The utility of the probabilistic response mode was examined first by comparing the usefulness of different scoring formulas associated with the response mode. Then, the factor structure resulting from the probabilistic response mode was studied in comparison to the factor structure obtained from scoring the responses dichotomously. Also examined in Research Report 83-3 were the validities of the scores obtained from the different scoring methods, their reliabilities, and the effects of certainty or risk-taking on the probabilistic scores.

Thompson's (1983) study also involved the administration of items in different response formats to college students. The study crossed two response formats (categorical and probabilistic) with two item types (multiple-choice and dichotomous) to obtain four different types of test items. These were (1) the conventional multiple-choice item; (2) a probabilistic multiple-choice item, similar to that used in Research Report 83-3; (3) a dichotomous (yes, no) item; and (4) a dichotomous-probabilistic item in which an examinee answered by stating, with a number between 0 and 100, his/her confidence that the answer to the question was the correct answer. Similar to Research Report 83-3, Thompson investigated several scoring systems for the probabilistic items. In addition, the four test item types were evaluated in terms of the intercorrelations of the scores they provided, their reliabilities, and their factor structures.

One other factor related to adaptive testing examined in this project concerned the effects of test administration variables on ability test performance and psychological reactions to testing. This study (Research Report 81-2) investigated the effects of two variables unique to computer administration. One variable was immediate knowledge of results of the correctness of each item re-

sponse during the process of test administration. The second variable--pacing of item presentation--was concerned with whether the pace of the test administration was controlled by the examinee or by the computer. The two variables were studied in both computer-administered conventional and adaptive tests. The dependent variables included ability test performance (maximum likelihood θ estimates and proportion correct), response pattern information, item response latencies, and psychological reactions to testing. Data were obtained from 477 college students who were randomly assigned to the experimental conditions.

Major Findings

Adaptive Achievement Testing

1. Adaptive intersubtest branching is a feasible approach to improving the efficiency of test administration when a test battery is adaptively administered. This approach can reduce test battery length by 50% or more with no appreciable effect on the psychometric characteristics of scores on the tests in the battery (Research Reports 79-6 and 81-4). Although the major reductions in test battery length were attributable to adaptive intrasubtest item selection, there were additional small reductions in test length due to intersubtest branching. Intersubtest branching also resulted in test battery information levels that closely approximated those of the full test battery, in comparison to information levels obtained solely from the use of adaptive intrasubtest item selection (Research Report 81-4). Results also indicated (Research Report 81-4) that the order in which subtests were selected for intersubtest branching had no effect on either the efficiency of test administration or on the psychometric characteristics of the resulting test scores.

2. The use of change scores to measure changes in achievement over time, which assumes that the factor underlying changes in performance is invariant, may be appropriate in some achievement testing environments and not in others. Results from college courses (Research Report 81-5) indicated that the factor structure of measured achievement in a biology course was not the same prior to instruction as it was after several weeks of instruction. In a mathematics course, however, the factor structure of measured achievement did not change over a 10-week period. These results suggest that in the absence of information to indicate that the dimensionality of measured achievement does not change over time, it is inappropriate to compute simple difference scores to measure changes in achievement levels.

3. There was some indication (Kingsbury, 1984) that Bayesian adaptive tests using an individual prior achievement level estimate resulted in more reliable change scores than were obtained from comparable conventional achievement tests. Further research is needed, however, to investigate the generalizability of these findings in other achievement domains.

4. Adaptive Mastery Testing (AMT) is a viable procedure for reducing test length of mastery tests and improving the efficiency of mastery classifications. In monte carlo simulation (Research Report 80-4) AMT achieved the best combination of test length reduction and validity of mastery classifications in comparison with a sequential probability ratio classification

procedure and conventional tests scored by proportion correct and IRT-based Bayesian scoring. The advantages of AMT were most pronounced in realistic item pools in which items varied in difficulties and discriminations. AMT also tended to result in a more even balance of false mastery and false non-mastery classifications in comparison to the sequential procedure.

Results from the simulation study were supported in live testing (Research Report 81-3). In comparison to conventional achievement tests, both fixed- and variable-length AMTs resulted in mastery classifications that were more consistent with an independent mastery criterion. The average variable-length adaptive test was able to make a high-confidence classification for students using only from 2 to 5 items, thus reducing test lengths as much as 74% to 88% from the 20-item conventional test, with no loss in classification accuracy.

Adaptive Ability Testing

5. Owen's Bayesian adaptive testing strategy results in θ estimates that, under realistic testing conditions, are biased and not of equal precision across θ levels (Research Report 83-2 and Weiss & McBride, 1984). Only under the unrealistic situation in which true θ was used as the prior θ did Owen's procedure result in unbiased θ estimates and reasonably horizontal information functions. Bias was also differentially affected by item discriminations for variable-length tests. In addition, for these tests, test length was an increasing function of θ. The design of these studies allowed identification of the source of the bias to be the use of a constant (group) prior θ estimate to begin the Bayesian adaptive testing.

6. Errors in item parameter estimates do not seriously affect the performance of adaptive testing strategies (Crichton, 1981; Mattson, 1983). In the 3-parameter data (Crichton, 1981) using indices combined across θ levels, when error was introduced into the separate item parameter estimates the effects were small for errors within the usually observed range. The a and b parameters generally had similar effects on adaptive test performance, while errors in the c parameter had negligible effects. When errors in the three parameters were combined, effects differed little from the case with error in a or b, except for very unrealistic levels of error. There were no appreciable differences in susceptibility to error among the stradaptive, maximum information, and Bayesian adaptive testing strategies.

7. When indices conditional on θ were examined in the 3-parameter data (Crichton, 1981), Bayesian and maximum information adaptive tests were somewhat less susceptible to errors in item parameter estimates than was the stradaptive test. Whereas errors in estimation of the b and c parameters had little effect on the conditional indices, estimation errors in the a parameter resulted in the major effects on the conditional indices, indicating that large errors in estimating a may deteriorate the performance of the adaptive testing strategies. Even with this deterioration in performance, however, the adaptive tests still performed better than the conventional tests for a substantial portion of the θ range.

8. When a and b parameter estimates were allowed to correlate with each other,

as they do in many real item pools, there was no additional effect on θ estimates beyond that due to errors in uncorrelated item parameter estimates (Mattson, 1983).

9. Maximum likelihood θ estimation performed better than Bayesian estimation for lesser degrees of error in the item parameters, and Bayesian estimation was less affected by item parameter erors for more extreme levels of error, particularly for the 1- and 2-parameter models (Mattson, 1983).

10. The 2-parameter model was least affected by errors in item parameter estimates (Mattson, 1983). Under conditions of large errors in item parameter estimates, the 2-parameter model performed better than the error-free cases of 1- and 3-parameter models.

11. Multidimensionality has a more serious effect on θ estimates from maximum information adaptive tests than does errors in item parameter estimates (Suhaldonik & Weiss, 1985). For multidimensional structures with one or two factors beyond the first that account for up to one-fourth the variance of the first factor, overcoming the effects of multidimensionality would require doubling of adaptive test length. The data also suggested that the number of factors, and not simply the overall strength of the factor structure, affects θ estimates, since a single factor beyond the first had less effect than did two factors that accounted for the same amount of variance. In general, however, adaptive testing is quite robust to multidimensional structures of the type most frequently resulting from careful item selection--i.e., factor structures with a strong first factor and second or third factors that account for less than one-eighth of the variance of the first factor.

12. Administration of multiple-choice items in a probabilistic response mode may be a useful application of computerized test administration. Although items answered in a probabilistic mode did not result in higher validities than multiple-choice items responded to dichotomously, the probabilistic mode resulted in higher reliabilities and a stronger first factor (Research Report 83-3). Since the stronger first factor would result in higher IRT item discrimination parameters for these items, adaptive testing based on items administered probabilistically would likely be more efficient, resulting in shorter tests or in more precise θ estimates. Additional analyses of item formats and response modes (Thompson, 1983) showed that items presented in a dichotomous format yielded different factor structures than did multiple-choice formats, but supported the higher reliabilities observed in Research Report 83-3 for the probabilistic response format.

13. Computerized test administration variables--including adaptive vs. conventional test type, computer- vs. self-paced item administration, and immediate knowledge of results after each item is administered--do not have direct effects on ability test performance, as measured by estimated θ levels (Research Report 81-2). These test administration variables do, however, have effects on psychological reactions to testing. Immediate knowledge of results appears to have a standardizing effect on test anxiety and test-taking motivation, since mean levels of anxiety and motivation were different when knowledge of results was provided but similar when it was not.

1 Dr. W. Steve Sellman
Office of the Assistant Secretary of Defense (MRA & L)
2B269 The Pentagon
Washington, DC 20301

1 Dr. Robert A. Wisher
OUSDRE (ELS)
The Pentagon, Room 3D129
Washington, DC 20301

Civilian Agencies

1 Dr. Patricia A. Butler
NIE Brn Bldg, Stop # 7
1200 19th St., NW
Washington, DC 20208

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 Mr. Thomas A. Warm
U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Private Sector

1 Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Tel Aviv, Ramat Aviv 69978
Israel

1 Dr. Werner Birke
Personalstammamt der Bundeswehr
D-5000 Koeln 90
WEST GERMANY

1 Dr. R. Darrell Bock
Department of Education
University of Chicago
Chicago, IL 60637

1 Mr. Arnold Bohrer
Section of Psychological Research
Caserne Petits Chateau
CRS
1000 Brussels
Belgium

1 Dr. Robert Brennan
American College Testing Programs
P. O. Box 168
Iowa City, IA 52243

1 Dr. Ernest R. Cadotte
307 Stokely
University of Tennessee
Knoxville, TN 37916

1 Dr. James Carlson
American College Testing Program
P.O. Box 168
Iowa City, IA 52243

1 Dr. John B. Carroll
409 Elliott Rd.
Chapel Hill, NC 27514

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

1 Dr. Hans Crombag
Education Research Center
University of Leyden
Boerhaavelaan 2
2334 EN Leyden
The NETHERLANDS

1 Lee Cronbach
16 Laburnum Road
Atherton, CA 94205

1 CTB/McGraw-Hill Library
2500 Garden Road
Monterey, CA 93940

1 Mr. Timothy Davey
University of Illinois
Department of Educational Psychology
Urbana, IL 61801

1 Dr. Dattpradad Divgi
Syracuse University
Department of Psychology
Syracuse, NE 33210

1 Dr. Emmanuel Donchin
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Dr. Hei-Ki Dong
Ball Foundation
800 Roosevelt Road
Building C, Suite 206
Glen Ellyn, IL 60137

1 Dr. Fritz Drasgow
Department of Psychology
University of Illinois
603 E. Daniel St.
Champaign, IL 61820

1 Dr. Stephen Dunbar
Lindquist Center for Measurement
University of Iowa
Iowa City, IA 52242

1 Dr. John M. Eddins
University of Illinois
252 Engineering Research Laboratory
103 South Mathews Street
Urbana, IL 61801

1 Dr. Susan Embertson
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
Lawrence, KS 66045

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

1 Dr. Benjamin A. Fairbank, Jr.
Performance Metrics, Inc.
5825 Callaghan
Suite 225
San Antonio, TX 78228

1 Dr. Leonard Feldt
Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

Marine Corps

1 Col. Ray Leidich
Headquarters, Marine Corps
MPI
Washington, DC 20380

1 Headquarters, U. S. Marine Corps
Code MPI-20
Washington, DC 20380

1 Special Assistant for Marine
Corps Matters
Code 100M
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

1 Major Frank Yohannan, USMC
Headquarters, Marine Corps
(Code MPI-20)
Washington, DC 20380

Army

1 Dr. Kent Eaton
Army Research Institute
5001 Eisenhower Blvd.
Alexandria, VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Clessen Martin
Army Research Institute
5001 Eisenhower Blvd.
Alexandria, VA 22333

1 Dr. Karen Mitchell
Army Research Institute
5001 Eisenhower Blvd
Alexandria, VA 22332

1 Dr. William E. Nordbrock
FMC-ADCO Box 25
APO, NY 09710

1 Dr. Harold F. O'Neil, Jr.
Director, Training Research Lab
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Commander, U.S. Army Research Institute
for the Behavioral & Social Sciences
ATTN: PERI-BR (Dr. Judith Orasanu)
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Mr. Robert Ross
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Robert Sasmor
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Joyce Shields
Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Hilda Wing
Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22333

Air Force

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Col. Roger Campbell
AF/MPXOA
Pentagon, Room 4E195
Washington, DC 20330

1 Mr. Raymond E. Christal
AFHRL/MOE
Brooks AFB, TX 78235

1 Dr. Alfred R. Fregly
AFOSR/NL
Bolling AFB, DC 20332

1 Dr. Patrick Kyllonen
AFHRL/MOE

Brooks AFB, TX 78235

1 Dr. Randolph Park
AFHRL/MOAN
Brooks AFB, TX 78235

1 Dr. Roger Pennell
Air Force Human Resources Laboratory
Lowry AFB, CO 80230

1 Dr. Malcolm Ree
AFHRL/MP
Brooks AFB, TX 78235

1 Maj. Bill Strickland
AF/MPXOA
4E168 Pentagon
Washington, DC 20330

1 Dr. John Tangney
AFOSR/NL
Bolling AFB, DC 20332

1 Major John Welsh
AFHRL/MOAN
Brooks AFB , TX 78223

1 Dr. Joseph Yasatuke
AFHRL/LRT
Lowry AFB, CO 80230

Department of Defense

12 Defense Technical Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
Attn: TC

1 Dr. Anita Lancaster
Accession Policy
OASD/MI&L/MP&FM/AP
Pentagon, Room 2B271
Washington, DC 20301

1 Dr. Jerry Lehnus
OASD (M&RA)
Washington , DC 20301

1 Dr. Clarence McCormick
HQ, MEPCOM
MEPCT-P
2500 Green Bay Road
Noprth Chicago, IL 60064

1 Military Assistant for Training and
Personnel Technology
Office of the Under Secretary of Defens
for Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

Distribution List

Navy

1 Code N711
Attn: Arthur S. Blaiwes
Naval Training Equipment Center
Orlando, FL 32813

Dr. Nick Bond
Office of Naval Research
Liaison Office, Far East
APO San Francisco, CA 96503

1 Lt. Alexander Bory
Applied Psychology
Measurement Division
NAMRL
NAS Pensacola, FL 32508

1 Dr. Robert Breaux
NAVTRAEQUIPCEN
Code N-095R
Orlando, FL 32813

1 Dr. Robert Carroll
NAVOP 115
Washington, DC 20370

1 Dr. Stanley Collyer
Office of Naval Technology
800 N. Quincy Street
Arlington, VA 22217

1 CDR Mike Curran
Office of Naval Research
800 N. Quincy St.
Code 270
Arlington, VA 22217

1 Dr. Charles E. Davis
Personnel and Training Research
Office of Naval Research (Code 442PT)
800 North Quincy Street
Arlington, VA 22217

1 Dr. John Ellis
Navy Personnel R&D Center
San Diego, CA 92252

1 DR. PAT FEDERICO
Code P13
NPRDC
San Diego, CA 92152

1 Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Ms. Rebecca Hetter
Navy Personnel R&D Center (Code 62)
San Diego, CA 92152

1 Mr. Dick Hoshaw
NAVOP-135
Arlington Annex
Room 2834
Washington, DC 20350

1 Dr. Norman J. Kerr
Chief of Naval Education and Training
Code 00A2
Naval Air Station
Pensacola, FL 32508

1 Dr. Leonard Kroeker
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Daryll Lang
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. William L. Maloy (02)
Chief of Naval Education and Training
Naval Air Station
Pensacola, FL 32508

1 Dr. James McBride
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr William Montague
NPRDC Code 13
San Diego, CA 92152

1 Ms. Kathleen Moreno
Navy Personnel R&D Center (Code 62)
San Diego, CA 92152

1 Library, Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

1 Technical Director
Navy Personnel R&D Center
San Diego, CA 92152

6 Personnel & Training Research Group
Code 442PT
Office of Naval Research
Arlington, VA 22217

1 Dr. Carl Ross
CNET-PDCD
Building 90
Great Lakes NTC, IL 60088

1 Mr. Drew Sands
NPRDC Code 62
San Diego, CA 92152

1 Dr. Mary Schratz
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Alfred F. Smode
Senior Scientist
Code 7B
Naval Training Equipment Center
Orlando, FL 32813

1 Dr. Richard Snow
Liaison Scientist
Office of Naval Research
Branch Office, London
Box 39
FPO New York, NY 09510

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

1 Mr. Brad Sympson
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Frank Vicino
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Ronald Weitzman
Naval Postgraduate School
Department of Administrative Sciences
Monterey, CA 93940

1 Dr. Douglas Wetzel
Code 12
Navy Personnel R&D Center
San Diego, CA 92152

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

1 Mr John H. Wolfe
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Wallace Wulfeck, III
Navy Personnel R&D Center
San Diego, CA 92152

Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Mattson, J. D. (1983, June). Effects of item parameter error and other factors on trait estimation in latent-trait-based adaptive testing. Unpublished doctoral dissertation, University of Minnesota.

Maurelli, V. A., & Weiss, D. J. (1981, November). Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries (Research Report 81-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Suhadolnik, D., & Weiss, D. J. (1983, July). Effect of examinee certainty on probabilistic test scores and a comparison of scoring methods for probabilistic responses (Research Report 83-3). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Suhadolnik, D., & Weiss, D. J. (1985). Robustness of adaptive testing to multidimensionality. In D. J. Weiss (Ed.), Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference (pp. 248-280). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Thompson, J. G. (1983, August). An investigation of the dimensionality of multiple-choice and dichotomous vocabulary test items administered in probabilistic and categorical response formats. Unpublished Master's thesis, University of Minnesota.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21, 361-375.

Weiss, D. J., & Davison, M. L. (1981, January). Review of test theory and methods (Research Report 81-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Weiss, D. J., & McBride, J. R. (1983, March). Bias and information of Bayesian adaptive testing (Research Report 83-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. Applied Psychological Measurement, 8, 273-285.

REFERENCES

Crichton, L. J. (1981, June). Effect of error in item parmeter estimates on adaptive testing. Unpublished doctoral dissertation, University of Minnesota.

Gialluca, K. A., & Weiss, D. J. (1979, November). Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement (Research Report 79-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Gialluca, K. A., & Weiss, D. J. (1981, December). Dimensionality of measured achievement over time (Research Report 81-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Johnson, M. F., Weiss, D. J., & Prestwood, J. S. (1981, February). Effects of immediate feedback and pacing of item presentation on ability test performance and psychological reactions to testing (Research Report 81-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Kingsbury, G. G. (1984, August). Adaptive self-referenced testing as a procedure for the measurement of individual change due to instruction: A comparison of the reliabilities of change estimates obtained from adaptive and conventional testing procedures. Unpublished doctoral dissertation, University of Minnesota.

Kingsbury, G. G., & Weiss, D. J. (1980a, September). A comparison of ICC-based adaptive mastery testing and the Waldian probability ratio method. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference (pp. 120-139). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Kingsbury, G. G., & Weiss, D. J. (1980b, November). A comparison of adaptive, sequential, and conventional testing strategies for mastery decisions (Research Report 80-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Kingsbury, G. G., & Weiss, D. J. (1980c, December). An alternate-forms reliability and concurrent validity comparison of Bayesian adaptive and conventional ability tests (Research Report 80-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Kingsbury, G. G., & Weiss, D. J. (1981, September). A validity comparison of adaptive and conventional strategies for mastery testing (Research Report 81-3). Minneapolis: University of Minnesota, Department of Psychology,

to each alternative as the item score. Total test scores for all of the scoring methods were obtained by summing individual item scores.

Several studies using probabilistic response methods have shown the effect of a response-style variable, called certainty or risk taking, on scores obtained from probabilistic responses. Results from this study showed a small effect of certainty on the probabilistic scores in terms of the validity of the scores but no effect at all on the factor structure or internal consistency of the scores. Once the effect of certainty on the probabilistic scores had been ruled out, the five scoring formulas were compared in terms of validity, reliability, and factor structure. There were no differences in the validity of the scores from the different methods, but scores obtained from the two scoring formulas that were not reproducing scoring systems were more reliable and had stronger first factors then the scores obtained using the reproducing scoring systems. For practical use, however, the reproducing scoring systems may have an advantage because they maximize examinees' scores when examinees respond honestly, while honest responses will not necessarily maximize an examinee's score with the other two methods. If a reproducing scoring system is used for this reason, the spherical scoring formula is recommended, since it was the most internally consistent and showed the strongest first factor of the reproducing scoring systems.

achievement levels increased the underlying factor structure remained unchanged. The implications of these results for psychology, education, and program evaluation are noted. (AD A110955)

Research Report 83-2
Bias and Information of Bayesian Adaptive Testing
David J. Weiss and James R. McBride
March 1983

Monte carlo simulation was used to investigate score bias and information characteristics of Owen's Bayesian adaptive testing strategy, and to examine possible causes of score bias. Factors investigated in three related studies included effects of an accurate prior θ estimate, effects of item discrimination, and effects of fixed vs. variable test length. Data were generated from a three-parameter logistic model for 3,100 simulees in each of eight data sets; Bayesian adaptive tests were administered, drawing items from a "perfect" item pool. Results showed that the Bayesian adaptive test yielded unbiased θ estimates and relatively flat information functions only in the unrealistic situation in which an accurate prior θ estimate was used. When a more realistic constant prior θ estimate was used with a fixed test length, severe bias was observed that varied with item discrimination. A different pattern of bias was observed with variable test length and a constant prior. Information curves for the constant prior conditions generally became more peaked and asymmetric with increasing item discrimination. In the variable test length condition the test length required to achieve a specified level of the posterior variance of θ estimates was an increasing function of θ level. These results indicate that θ estimates from Owen's Bayesian adaptive testing method are affected by the prior θ estimate used and that the method does not provide measurements that are unbiased and equiprecise except under the unrealistic condition of an accurate prior θ estimate. (AD A129280)

Research Report 83-3
Effect of Examinee Certainty on Probabilistic Test Scores
and a Comparison of Scoring Methods for Probabilistic Responses
Debra Suhadolnik and David J. Weiss
July 1983

The present study was an attempt to alleviate some of the difficulties inherent in multiple-choice items by having examinees respond to multiple-choice items in a probabilistic manner. Using this format, examinees are able to respond to each alternative and to provide indications of any partial knowledge they may possess concerning the item. The items used in this study were 30 multiple-choice analogy items. Examinees were asked to distribute 100 points among the four alternatives for each item according to how confident they were that each alternative was the correct answer. Each item was scored using five different scoring formulas. Three of these scoring formulas--the spherical, quadratic, and truncated log scoring methods--were reproducing scoring systems. The fourth scoring method used the probability assigned to the correct alternative as the item score, and the fifth used a function of the absolute difference between the correct response vector for the four alternatives and the actual points assigned

Research Report 81-4
Factors Influencing the Psychometric Characteristics of an Adaptive Testing Strategy for Test Batteries
Vincent A. Maurelli and David J. Weiss
November 1981

A monte carlo simulation was conducted to assess the effects in an adaptive testing strategy for test batteries of varying subtest order, subtest termination criterion, and variable versus fixed entry on the psychometric properties of an existent achievement test battery. Comparisons were made among conventionally administered tests and adaptive tests using adaptive intra-subtest item selection with and without inter-subtest branching. Data consisted of responses of 300 simulees to a 201-item achievement test battery. Mean test battery length was reduced from 42.5% to 52.3% using adaptive intra-subtest item selection with variable termination. Reductions in mean subtest lengths ranged from 27% to 67%. When inter-subtest branching was added, additional test length reductions of 1% to 2% were observed for individual subtests. The reductions in test length were achieved with no significant loss of fidelity or psychometric information. The addition of inter-subtest branching resulted in levels of mean test battery information more similar to those of the full test battery, even with mean test battery reductions of 50% in number of items administered. Subtest order was shown to have no effect on the evaluative criteria employed. The results generally supported previous studies of this adaptive testing strategy. Suggestions for future research are presented. (AD A109666)

Research Report 81-5
Dimensionality of Measured Achievement Over Time
Kathleen A. Gialluca and David J. Weiss
December 1981

Some type of difference or change score is frequently used to quantify the effects of experimental treatments and educational programs on individuals and on groups of individuals. Whether the change measurement involves the use of simple difference scores, their derivatives, or some more complex methodological design, the measurement process itself assumes that the treatment or instruction results in higher levels of the originally measured variable and that the only change that occurs is a quantitative one. If this assumption is not met, then the computation of any type of difference score is inappropriate and the scores themselves are useless for measuring growth or change.

Two studies investigated the tenability of the assumption that classroom instruction results in increases in students' achievement levels while the qualitative nature of that achievement remains constant across time. The data utilized were the item responses to tests in basic mathematics and in general biology administered as pretests and after instruction to students enrolled in those courses.

Results indicated that this assumption was not tenable in the biology data set, where increases in mean achievement level were accompanied by corresponding changes in the factor structure underlying the item responses. For the mathematics data, however, there was no such violation of the assumption: As student

These results indicate that testing conditions may interact in a complex way to determine psychological reactions to the testing environment. The interactions do suggest, however, a somewhat consistent standardizing effect of KR on test anxiety and test-taking motivation. This standardizing effect of KR showed that approximately equal levels of motivation and anxiety were reported under the various testing conditions when KR was provided, but that mean levels of these variables were substantially different when KR was not provided. Consistent with theoretical expectations, the conventional test was perceived as being either too easy or too difficult, whereas the adaptive tests were perceived more often as being of appropriate difficulty.

The results concerning the effects of KR on test performance, motivation, and anxiety found in this study were contrary to earlier reported findings; and differences in the studies are delineated. Recommendations are made concerning the control of specific testing conditions, such as difficulty of the test and ability level of the examinee population, as well as suggestions for the further analysis of the standardizing effect of KR. (AD A097688)

Research Report 81-3
A Validity Comparison of Adaptive and Conventional Strategies for Mastery Testing
G. Gage Kingsbury and David J. Weiss
September 1981

Conventional mastery tests designed to make optimal mastery classifications were compared with fixed-length and variable-length adaptive mastery tests in terms of validity of decisions with respect to an external criterion measure. Comparisons between the testing procedures were made across five content areas in an introductory biology course from tests administered to over 400 volunteer students. The criterion measure used was the student's final standing in the course, based on course examinations and laboratory grades. Results indicated that the adaptive test resulted in mastery classifications that were more consistent with final class standing than those obtained from the conventional test. This result was observed within individual content areas and for discriminant analysis classifications made across content areas. This result was also observed for two scoring procedures used with the conventional test (proportion-correct and Bayesian scoring). Results also indicated that there was no decrement in the performance of the adaptive test when a variable termination rule was implemented. This variable termination rule resulted in test lengths which were, on the average, 74% to 88% shorter than the original adaptive tests. Further analyses explicated the manner in which the adaptive tests administered differed from the conventional test for each content area as a function of achievement level. This evidence was used to explain why the adaptive tests resulted in more valid decisions than the conventional procedure, in spite of the fact that the type of conventional test used here was the most informative test concerning the mastery cutoff. It is concluded that variable-length adaptive mastery tests can provide more valid mastery classifications than "optimal" conventional mastery tests while reducing test length an average of 80% from the length of the conventional tests. (AD A106867)

plied to problems of item option weighting and adaptive testing. Important developments with these models during the period included the demonstration of their relationship with other psychological measurement models, and methods for determining fit of individuals to IRT models. As another alternative to classical test theory, order models were developed and studied, and several other models were proposed.

Validity issues were also studied during this period. A number of approaches to the analysis of multitrait-multimethod matrices were proposed and compared, including some based on structural equations models. Issues of predictive validity studied included necessary sample sizes, validity generalization, and moderator and suppressor effects. Test fairness issues and their effects on validity received considerable attention. Concern was with (1) bias in selection; (2) fairness to minorities, including differential and single-groups validity and comparisons of regression lines, adverse impact, and bias in test content; and (3) fairness to women.

It is concluded that little of consequence was accomplished in classical test theory during this period. The most important developments were in alternatives to classical test theory, primarily item response theory. Research in this area resulted in data and other developments that will permit a better understanding of the range of applicability of these models and their potential for solving measurement problems not solvable by classical models. (AD A096157)

Research Report 81-2
Effects of Immediate Feedback and Pacing of Item Presentation
on Ability Test Performance and Psychological Reactions to Testing
Marilyn F. Johnson, David J. Weiss, and J. Stephen Prestwood
February 1981

The study investigated the joint effects of knowledge of results (KR or no-KR), pacing of item presentation (computer or self-pacing), and type of testing strategy (50-item peaked conventional, variable-length stradaptive, or 50-item fixed-length stradaptive test) on ability test performance, test item response latency, information, and psychological reactions to testing. The psychological reactions to testing were obtained from Likert-type items that assessed test-taking anxiety, motivation, perception of difficulty, and reactions to knowledge of results. Data were obtained from 447 college students randomly assigned to one of the 12 experimental conditions.

The results indicated that there were no effects on ability estimates due to knowledge of results, testing strategy, or pacing of item presentation. Although average latencies were greater on the stradaptive tests than on the conventional test, the overall testing time was not substantially longer on the adaptive tests and may have been a function of differences in test difficulty. Analysis of information values indicated higher levels of information on the stradaptive tests than on the conventional test. There was no statistically significant main effect for any of the three experimental conditions when test anxiety or test-taking motivation were the dependent variables, although there were some significant interaction effects.

The concurrent validity analysis showed that the conventional test produced ability level estimates that correlated more highly with the criterion test scores than did the Bayesian test for all lengths greater than four items. This result was observed for both scoring procedures used with the conventional test.

Limitations of the study, and the conclusions that may be drawn from it, are discussed. These limitations, which may have affected the results of this study, included possible differences in the alternate forms used within the two testing strategies, the relatively small calibration samples used to estimate the ICC parameters for the items used in the study, and method variance in the conventional tests. (AD A094477)

Research Report 81-1
Review of Test Theory and Methods
David J. Weiss and Mark L. Davison
January 1981

The research literature on test theory and methods for the period 1975 through early 1980 is critically reviewed. Research on classical test theory has concentrated on relatively unimportant developments in reliability theory, with some new developments and applications of generalizability theory appearing during this period. The reliability of change or gain scores has received some attention from the classical test theory perspective, as have the applications of classical reliability concepts in experimental design and the analysis of experimental data. A minor amount of research with classical models was in the area of test-score equating. Classical item analysis procedures, however, received little attention. A fair amount of research during the period was devoted to different item types and test item response modes as replacements for the ubiquitous multiple-choice item. Several types of true-false items were proposed, and formula scoring was studied by a number of researchers in an attempt to reduce guessing effects. The perennial topic of response option weighting received attention, with efforts oriented toward demonstrating effects on validity and reliability. Response modes studied included answer-until-correct, confidence weighting, and free-response.

A number of alternatives to classical test theory were studied in an attempt to solve some of the problems for which classical test theory has proven to be inadequate. Research on criterion-referenced testing continued during this period. Latent trait test theory (item response theory, or IRT) received considerable attention. Research on the 1-parameter IRT model continued to address problems of parameter estimation, model fit, and equating. The question of the person-free and sample-free characteristics of this model (i.e., its robustness) were investigated, with results generally supporting these desirable characteristics. In addition, a special case of this model that can account for guessing was developed, and the model was generalized and successfully applied to polychotomous attitude types of items. Considerable research occurred on the 2- and 3-parameter IRT models. The concept of information as a replacement for classical reliability concepts was studied, and its uses in developing parallel tests were described. As with the 1-parameter IRT model, problems of parameter estimation and equating were investigated. These IRT models were successfully ap-

Research Report 80-4
A Comparison of Adaptive, Sequential, and Conventional Testing Strategies for Mastery Decisions
G. Gage Kingsbury and David J. Weiss
November 1980

Two procedures for making mastery decisions with variable length tests and a conventional mastery testing procedure were compared in monte carlo simulation. The simulation varied the characteristics of the item pool used for testing and the maximum test length allowed. The procedures were compared in terms of the mean test length needed to make a decision, the validity of the decisions made by each procedure, and the types of classification errors made by each procedure. Both of the variable test length procedures were found to result in important reductions in mean test length from the conventional test length. The Sequential Probability Ratio Test (SPRT) procedure resulted in greater test length reductions, on the average, than the Adaptive Mastery Testing (AMT) procedure. However, the AMT procedure resulted both in more valid mastery decisions and in more balanced error rates than the SPRT procedure under all conditions. In addition, the AMT procedure produced the best combination of test length and validity. (AD A094478)

Research Report 80-5
An Alternate-Forms Reliability and Concurrent Validity Comparison of Bayesian Adaptive and Conventional Ability Tests
G. Gage Kingsbury and David J. Weiss
December 1980

Two 30-item alternate forms of a conventional test and a Bayesian adaptive test were administered by computer to 472 undergraduate psychology students. In addition, each student completed a 120-item paper-and-pencil test, which served as a concurrent validity criterion test, and a series of very easy questions designed to detect students who were not answering conscientiously. All test items were five-alternative multiple-choice vocabulary items. Reliability and concurrent validity of the two testing strategies were evaluated after the administration of each item for each of the tests, so that trends indicating differences in the testing strategies as a function of test length could be detected. For each test, additional analyses were conducted to determine whether the two forms of the test were operationally alternate forms.

Results of the analysis of alternate-forms correspondence indicated that for all test lengths greater than 10 items, each of the alternate forms for the two test types resulted in fairly constant mean ability level estimates. When the scoring procedure was equated, the mean ability levels estimated from the two forms of the conventional test differed to a greater extent than those estimated from the two forms of the Bayesian adaptive test.

The alternate-forms reliability analysis indicated that the two forms of the Bayesian test resulted in more reliable scores than the two forms of the conventional test for all test lengths greater than two items. This result was observed when the conventional test was scored either by the Bayesian or proportion-correct method.

ABSTRACTS OF RESEARCH REPORTS

Research Report 79-6
Efficiency of an Adaptive Inter-Subtest Branching Strategy
in the Measurement of Classroom Achievement
Kathleen A. Gialluca and David J. Weiss
November 1979

A real-data simulation was conducted to investigate the efficiency of an adaptive testing strategy designed for achievement test batteries applied to a classroom achievement test. This testing strategy combined adaptive item selection routines both within and between the subtests of the test battery. Comparisons were made between the conventionally-administered tests and the simulated adaptive tests in terms of test length, psychometric information, and correlations of achievement estimates. Design of the study also permitted (1) separation of the effects of the adaptive intra-subtest item selection procedure and inter-subtest branching, (2) evaluation of the effects of different intra-subtest termination criteria, (3) use of classical regression equations and regression equations corrected for errors of measurement in the predictors, and (4) cross-validation stability of the inter-subtest branching regression predictions. Data consisted of the responses from 1,600 students to classroom-administered final exams in a general biology course at the University of Minnesota.

Total test length was reduced from 16% to 30% using the adaptive intra-subtest item selection strategy with a variable termination criterion that omits those items providing little information to the measurement process. Subtest-length reductions ranged from about 8% to 62%. Total test length was reduced another 1% to 5% (with subtest-length reductions of up to 53%) upon the addition of an inter-subtest branching strategy that utilized regression equations with prior information concerning a student's performance.

Reductions in subtest length were accomplished with virtually no loss in psychometric information. Correlations between the Bayesian achievement estimates from the adaptive and conventional tests were uniformly high, typically $\underline{r} = .90$ and higher. Results showed that the use of the corrected regression equations did little to improve the performance of the inter-subtest branching; although the multiple correlations for the corrected equations were higher, both the information curves and correlations of achievement estimates were generally lower. Cross-validation results indicated that the procedure can be used in different samples from the same population.

Results from this study generally supported the generality of this adaptive testing strategy for reducing achievement test length with no adverse impact on the quality of the measurements. Suggestions are made for further research with this testing strategy. (AD A080956)

Perceptions of test difficulty were different for adaptive and conventional tests; students accurately perceived the conventional tests as either too easy or too difficult, depending on their ability levels, while the adaptive test was generally accurately perceived as being of appropriate difficulty.

1 Univ. Prof. Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

1 Professor Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

1 Dr. Dexter Fletcher
University of Oregon
Department of Computer Science
Eugene, OR 97403

1 Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01002

1 Dr. Robert Glaser
Learning Research & Development Center
University of Pittsburgh
3939 O'Hara Street
PITTSBURGH, PA 15260

1 Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

1 Dipl. Pad. Michael W. Habon
Universitat Dusseldorf
Erziehungswissenschaftliches Inst. II
Universitatsstr. 1
D-4000 Dusseldorf 1
WEST GERMANY

1 Dr. Ron Hambleton
School of Education
University of Massachusetts
Amherst, MA 01002

1 Dr. Delwyn Harnisch
University of Illinois
51 Gerty Drive
Champaign, IL 61820

1 Prof. Lutz F. Hornke
Universitat Dusseldorf
Erziehungswissenschaftliches Inst. II
Universitatsstr. 1
Dusseldorf 1
WEST GERMANY

1 Dr. Paul Horst
677 G Street, #184
Chula Vista, CA 90010

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
603 East Daniel Street
Champaign, IL 61820

1 Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

1 Dr. Jack Hunter
2122 Coolidge St.
Lansing, MI 48906

1 Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208

1 Dr. Douglas H. Jones
Advanced Statistical Technologies
Corporation
10 Trafalgar Court
Lawrenceville, NJ 08148

1 Professor John A. Keats
Department of Psychology
The University of Newcastle
N.S.W. 2308
AUSTRALIA

1 Dr. William Koch
University of Texas-Austin
Measurement and Evaluation Center
Austin, TX 78703

1 Dr. Thomas Leonard
University of Wisconsin
Department of Statistics
1210 West Dayton Street
Madison, WI 53705

1 Dr. Alan Lesgold
Learning R&D Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

1 Dr. Michael Levine
Department of Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

1 Dr. Charles Lewis
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat 23
9712GC Groningen
Netherlands

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

1 Dr. Robert Lockman
Center for Naval Analysis
200 North Beauregard St.
Alexandria, VA 22311

1 Dr. Frederic M. Lord
Educational Testing Service
Princeton, NJ 08541

1 Dr. James Lumsden
Department of Psychology
University of Western Australia
Nedlands W.A. 6009
AUSTRALIA

1 Dr. Gary Marco
Stop 31-E
Educational Testing Service
Princeton, NJ 08451

1 Mr. Robert McKinley
University of Toledo
Dept of Educational Psychology
Toledo, OH 43606

1 Dr. Barbara Means
Human Resources Research Organization
300 North Washington
Alexandria, VA 22314

1 Dr. Robert Mislevy
Educational Testing Service
Princeton, NJ 08541

1 Dr. W. Alan Nicewander
University of Oklahoma
Department of Psychology
Oklahoma City, OK 73069

1 Dr. Melvin R. Novick
356 Lindquist Center for Measurement
University of Iowa
Iowa City, IA 52242

1 Dr. James Olsen
WICAT, Inc.
1875 South State Street
Orem, UT 84057

1 Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupont Circle, NW
Washington, DC 20036

1 Dr. James Paulson
Dept. of Psychology
Portland State University
P.O. Box 751
Portland, OR 97207

1 Dr. Mark D. Reckase
ACT
P. O. Box 168
Iowa City, IA 52243

1 Dr. Lawrence Rudner
400 Elm Avenue
Takoma Park, MD 20012

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

1 PROF. FUMIKO SAMEJIMA
DEPT. OF PSYCHOLOGY
UNIVERSITY OF TENNESSEE
KNOXVILLE, TN 37916

1 Frank L. Schmidt
Department of Psychology
Bldg. 66
George Washington University
Washington, DC 20052

1 Lowell Schoer
Psychological & Quantitative
Foundations
College of Education
University of Iowa
Iowa City, IA 52242

1 Dr. Kazuo Shigemasu
7-9-24 Kugenuma-Kaigan
Fujisawa 251
JAPAN

1 Dr. William Stout
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

1 Dr. H. Wallace Sinaiko
Program Director
Manpower Research and Advisory Services
Smithsonian Institution
801 North Pitt Street
Alexandria, VA 22314

1 Martha Stocking
Educational Testing Service
Princeton, NJ 08541

1 Dr. Peter Stoloff
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

1 Dr. William Stout
University of Illinois
Department of Mathematics
Urbana, IL 61801

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and
Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Kikumi Tatsuoka
Computer Based Education Research Lab
252 Engineering Research Laboratory
Urbana, IL 61801

1 Dr. Maurice Tatsuoka
220 Education Bldg
1310 S. Sixth St.
Champaign, IL 61820

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Mr. Gary Thomasson
University of Illinois
Department of Educational Psychology
Champaign, IL 61829

1 Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

1 Dr. Ledyard Tucker
University of Illinois
Department of Psychology
603 E. Daniel Street
Champaign, IL 61820

1 Dr. David Vale
Assessment Systems Corporation
2233 University Avenue
Suite 310
St. Paul, MN 55114

1 Dr. Howard Wainer
Division of Psychological Studies
Educational Testing Service
Princeton, NJ 08540

1 Dr. Ming-Mei Wang
Lindquist Center for Measurement
University of Iowa
Iowa City , IA 52242

1 Dr. Brian Waters
HumRRO
300 North Washington
Alexandria, VA 22314

1 Dr. Rand R. Wilcox
University of Southern California
Department of Psychology
Los Angeles, CA 90007

1 German Military Representative
ATTN: Wolfgang Wildegrube
Streitkraefteamt
D-5300 Bonn 2
4000 Brandywine Street, NW
Washington , DC 20016

1 Dr. Bruce Williams
Department of Educational Psychology
University of Illinois
Urbana, IL 61801

1 Ms. Marilyn Wingersky
Educational Testing Service
Princeton, NJ 08541

1 Dr. George Wong
Biostatistics Laboratory
Memorial Sloan-Kettering Cancer Center
1275 York Avenue
New York, NY 10021

1 Dr. Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

Previous Publications (continued)

78-3. A Comparison of Levels and Dimensions of Performance in Black and White Groups
78-2. The Effects of Knowledge of Results and Test Difficulty on Ability Test Performance and Psychological Reactions to Testing. September 1978.
78-1. A Comparison of the Fairness of Adaptive and Conventional Testing Strategies. August 1978.
77-7. An Information Comparison of Conventional and Adaptive Tests in the Measurement of Classroom Achievement. October 1977.
77-6. An Adaptive Testing Strategy for Achievement Test Batteries. October 1977.
77-5. Calibration of an Item Pool for the Adaptive Measurement of Achievement. September 1977.
77-4. A Rapid Item-Search Procedure for Bayesian Adaptive Testing. May 1977.
77-3. Accuracy of Perceived Test-Item Difficulties. May 1977.
77-2. A Comparison of Information Functions of Multiple-Choice and Free-Response Vocabulary Items. April 1977.
77-1. Applications of Computerized Adaptive Testing. March 1977.
Final Report: Computerized Ability Testing, 1972-1975. April 1976.
76-5. Effects of Item Characteristics on Test Fairness. December 1976.
76-4. Psychological Effects of Immediate Knowledge of Results and Adaptive Ability Testing. June 1976.
76-3. Effects of Immediate Knowledge of Results and Adaptive Testing on Ability Test Performance. June 1976.
76-2. Effects of Time Limits on Test-Taking Behavior. April 1976.
76-1. Some Properties of a Bayesian Adaptive Ability Testing Strategy. March 1976.
75-6. A Simulation Study of Stradaptive Ability Testing. December 1975.
75-5. Computerized Adaptive Trait Measurement: Problems and Prospects. November 1975.
75-4. A Study of Computer-Administered Stradaptive Ability Testing. October 1975.
75-3. Empirical and Simulation Studies of Flexilevel Ability Testing. July 1975.
75-2. TETREST: A FORTRAN IV Program for Calculating Tetrachoric Correlations. March 1975.
75-1. An Empirical Comparison of Two-Stage and Pyramidal Adaptive Ability Testing. February 1975.
74-5. Strategies of Adaptive Ability Measurement. December 1974.
74-4. Simulation Studies of Two-Stage Ability Testing. October 1974.
74-3. An Empirical Investigation of Computer-Administered Pyramidal Ability Testing. July 1974.
74-2. A Word Knowledge Item Pool for Adaptive Ability Measurement. June 1974.
74-1. A Computer Software System for Adaptive Ability Measurement. January 1974.
73-4. An Empirical Study of Computer-Administered Two-Stage Ability Testing. October 1973.
73-3. The Stratified Adaptive Computerized Ability Test. September 1973.
73-2. Comparison of Four Empirical Item Scoring Procedures. August 1973.
73-1. Ability Measurement: Conventional or Adaptive? February 1973.

Copies of these reports are available, while supplies last, from:

Computerized Adaptive Testing Laboratory
N660 Elliott Hall
University of Minnesota
75 East River Road
Minneapolis MN 55455 U.S.A.

Previous Publications

Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference. March 1985.

Proceedings of the 1979 Computerized Adaptive Testing Conference. September 1980

Proceedings of the 1977 Computerized Adaptive Testing Conference. July 1978.

Research Reports

83-3. Effect of Examinee Certainty on Probabilistic Test Scores and a Comparison of Scoring Methods for Probabilistic Responses. July 1983.

83-2. Bias and Information of Bayesian Adaptive Testing. March 1983.

83-1. Reliability and Validity of Adaptive and Conventional Tests in a Military Recruit Population. January 1983.

81-5. Dimensionality of Measured Achievement Over Time. December 1981.

81-4. Factors Influencing the Psychometric Characteristics of an Adaptive Testing Strategy for Test Batteries. November 1981.

81-3. A Validity Comparison of Adaptive and Conventional Strategies for Mastery Testing. September 1981.

Final Report: Computerized Adaptive Ability Testing. April 1981.

81-2. Effects of Immediate Feedback and Pacing of Item Presentation on Ability Test Performance and Psychological Reactions to Testing. February 1981.

81-1. Review of Test Theory and Methods. January 1981.

80-5. An Alternate-Forms Reliability and Concurrent Validity Comparison of Bayesian Adaptive and Conventional Ability Tests. December 1980.

80-4. A Comparison of Adaptive, Sequential, and Conventional Testing Strategies for Mastery Decisions. November 1980.

80-3. Criterion-Related Validity of Adaptive Testing Strategies. June 1980.

80-2. Interactive Computer Administration of a Spatial Reasoning Test. April 1980.

Final Report: Computerized Adaptive Performance Evaluation. February 1980.

80-1. Effects of Immediate Knowledge of Results on Achievement Test Performance and Test Dimensionality. January 1980.

79-7. The Person Response Curve: Fit of Individuals to Item Characteristic Curve Models. December 1979.

79-6. Efficiency of an Adaptive Inter-Subtest Branching Strategy in the Measurement of Classroom Achievement. November 1979.

79-5. An Adaptive Testing Strategy for Mastery Decisions. September 1979.

79-4. Effect of Point-in-Time in Instruction on the Measurement of Achievement. August 1979.

79-3. Relationships among Achievement Level Estimates from Three Item Characteristic Curve Scoring Methods. April 1979.

Final Report: Bias-Free Computerized Testing. March 1979.

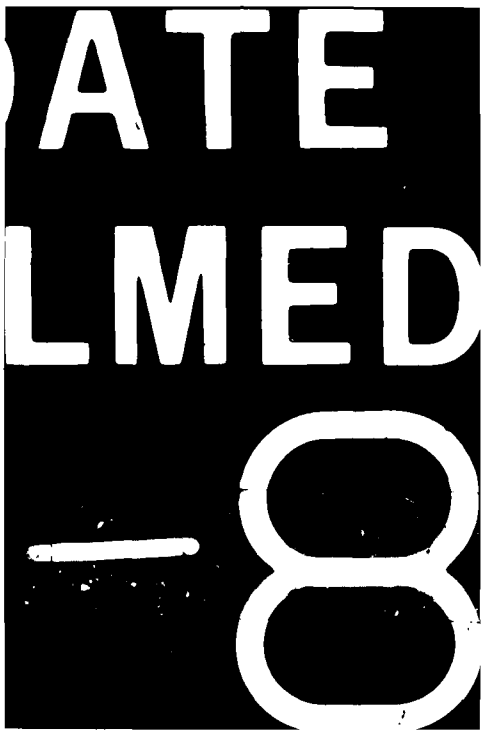
79-2. Effects of Computerized Adaptive Testing on Black and White Students. March 1979.

79-1. Computer Programs for Scoring Test Data with Item Characteristic Curve Models. February 1979.

78-5. An Item Bias Investigation of a Standardized Aptitude Test. December 1978.

78-4. A Construct Validation of Adaptive Achievement Testing. November 1978. on Tests of Vocabulary, Mathematics, and Spatial Ability. October 1978.

-continued inside-


ATE
LMED
8